Uncalibrated and Unsynchronized Human Motion Capture : A Stereo Factorization Approach

P. Tresadern and I. Reid,
Dept. of Engineering Science, University of Oxford,
Oxford, England OX1 3PJ
{pat,ian}@robots.ox.ac.uk

Abstract

Human motion capture typically requires several high quality, synchronized and calibrated cameras in a studio environment and can be potentially costly and technically complex. Instead, we propose a system which combines and improves upon two existing techniques, yielding an efficient method that recovers maximum likelihood joint angles and anthropomorphic data of the subject by factorization.

The first technique concerns using a rank constraint framework to synchronize sequences of non-rigid motions where we extend affine methods to perspective and homography projection models. The second is a self-calibration method for two affine cameras, using constraints derived from prior knowledge of the underlying structure. We propose a minimal parameterization of the system to obtain an initial solution then apply a full bundle adjustment over the free parameters based on a geometric error.

We demonstrate the efficacy of our method by comparing the recovered structure and motion with that from a commercial motion capture system.

1. Introduction

Commercial human motion capture systems [20] use multiple hardware-synchronized, accurately calibrated cameras under controlled conditions to track high contrast markers at the joints, reconstructing the pose of the subject by triangulation[1]. Markerless motion capture systems proposed by the research community (e.g. [1, 6, 7, 10]) use fewer cameras but often retain the requirement of accurate calibration and universally assume the cameras to be synchronized.

In contrast, we propose a system using only two uncalibrated and unsynchronized cameras. Not only does this reduce cost and complexity but could also be employed for surveillance or sporting analysis where the calibration and synchronization of the cameras is unavailable. Our system combines and develops methods for synchronizing image sequences with those for stereo self-calibration (also referred to as 'rectification') using non-rigid affine structure.

We extend current rank constraint based methods for the affine projection model [21, 19] to perspective and homography models. In contrast to [4, 12] we use an algebraic (rather than geometric) distance measure that is computationally inexpensive and can be expressed in a rank constraint framework. For self-calibration we exploit a minimal parameterization of a matrix $\mathbf{\Omega}$ that reduces computational complexity, guarantees an intuitive initialization, implicitly enforces 'positive definiteness' and requires no empirically 'tuned' parameters (see Section 2.2). Finally, we apply a bundle adjustment over all free parameters to correctly minimize a geometric (rather than algebraic) reprojection error.

We review existing methods in Section 2 before outlining our extensions to synchronization methods in Section 3. The improved self-calibration procedure is presented in Section 4 and the complete system on a novel sequence in Section 5. Section 6 concludes and discusses future work.

2. Related work

Our work is based upon the principle of factorization introduced by Tomasi and Kanade [17] who showed that, under orthographic projection, the rank of a normalized $2V \times N$ 'measurement matrix', $\mathbf{W}$, is bounded above by 3 and that the best[2] estimate of structure and motion is achieved by factorizing $\mathbf{W}$ using the Singular Value Decomposition (SVD) and discarding all data relating to the singular values $\sigma_4, \ldots, \sigma_r$.

In the context of dynamic scene stereoposis, $\mathbf{W}(f, f')$ is computed using measurements from frames f and f' of the 'reference' and 'target' sequences, respectively:

$$\mathbf{W}(f,f') = \begin{bmatrix} \mathbf{x}^{f}_{ref,1} & \cdots & \mathbf{x}^{f}_{ref,N} \\ \mathbf{x}^{f'}_{tgt,1} & \cdots & \mathbf{x}^{f'}_{tgt,N} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{bmatrix} \mathbf{X}^f \quad (1)$$

[1] Using markers solves the spatial correspondence problem which we do not explicitly address here. Instead, we assume spatial correspondence is available from markers or via manual labelling of the input images.

[2] Reid and Murray [14] later qualified 'best' by showing that the computed structure and motion minimizes reprojection error and can therefore be interpreted as a Maximum Likelihood estimate, assuming isotropic Gaussian noise.

where $\mathbf{x}^{f}_{i,\,n}$ denotes the image coordinates of the nth feature in the fth frame of the ith sequence, $\mathbf{P}_i$ is the ith projection matrix and $\mathbf{X}^f$ is the computed structure at frame f. This method was later extended for multiple objects by Costeira and Kanade [5] although the application of this extension has not yet been exploited for articulated motions (in part due to its failure in the presence of dependencies between motions [22]).

2.1. Synchronization

When presented with two sequences captured from unsynchronized cameras (Figure 1), we wish to recover corresponding frames related by the 1D affine transformation $f' = \alpha f + \delta t$ where α is the frame rate ratio and δt is the offset between the 0th frames in each sequence.

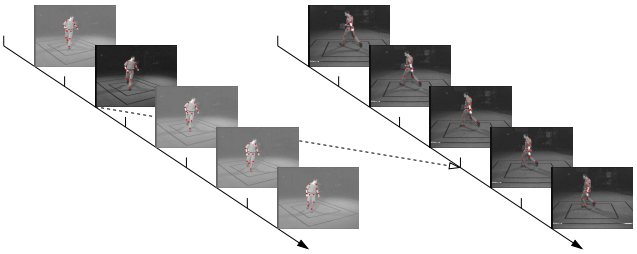

Figure 1: Two timelines with synchronization shown by the arrow. Note that a frame may not physically exist at the corresponding target instant due to the cameras being unsynchronized.

The alignment of image sequences has been studied by Caspi and Irani recovering the spatiotemporal alignment of sequences using optical flow with a planar projective (homography) model [2, 3]. However, our work is more closely related to their feature-based methods for wide baseline stereo [4]. Forming putative matches between feature *tracks* and utilizing a voting scheme they compute the spatiotemporal relationship between the views using the geometric distance between points and their associated epipolar lines in the manner suggested by Reid and Zisserman [15].

The same error metric is also used by Pooley *et al* [12] who compute the epipolar geometry of the cameras using matched background features and estimate synchronization parameters using the Hough transform. Zhou and Tao [23] assume a linear trajectory over a small period of time and use the epipolar geometry for feature transfer over two consecutive frames to estimate the temporal offset from the cross ratio of sets of four points.

In contrast, our work is inspired by that of Wolf and Zomet [21] who exploit rank constraints to recover corresponding frames by minimizing the 'energy', g, of a rank R matrix above its *expected* rank bound, r, defined as the sum of squared singular values $g = \sum_{i=r+1}^{R} \sigma_i^2$. This was later extended [19] to recover synchronization to *sub-frame* accuracy for sequences of *unknown and differing frame rates*.

It can be shown [18] that the 'energy' of $\mathbf{W}(f, f')$ as defined by Wolf and Zomet is equal to the reprojection error following factorization and is therefore an intuitively appropriate metric for determining synchronization - at alignment the image features are consistent with an underlying interpretation of three-dimensional structure (the pose at that instant), whereas if the sequences are not aligned the images are of *different* points in space and therefore not subject to the rank constraint.

2.2. Metric rectification

Given matched feature coordinates from multiple views, affine structure and motion can be recovered from the factorization of $\mathbf{W}$:

$$\mathbf{W} = \mathbf{P}\mathbf{X} = \mathbf{P}\mathbf{A}^{-1}\mathbf{A}\mathbf{X} \tag{2}$$

where $\mathbf{A}$ is a 3×3 matrix[3]. The QR-decomposition of $\mathbf{A} \rightarrow \mathbf{S}\mathbf{U}$ gives a 3D rotation, $\mathbf{S}$, and a pure affine transformation, $\mathbf{U}$. Since $\mathbf{S}$ simply effects a change of Euclidean coordinate frame *after rectification* it can be discarded without loss of generality. Therefore, to recover joint angles (which are not preserved in an affine coordinate frame) we must determine the rectifying affine transformation, $\mathbf{U}$. We define a matrix $\mathbf{\Omega} = \mathbf{U}^T\mathbf{U}$ such that $\mathbf{U}$ is recovered by Cholesky factorization if and only if $\mathbf{\Omega}$ is positive definite.

In contrast to classical calibration methods [9] that utilize known world distances, it was shown [17] that constraints on the recovered *motion* are sufficient to perform self-calibration, as formalized by Quan for all parallel projection models [13]. Specifically, the axes $\mathbf{i}^T$ and $\mathbf{j}^T$ transform to $\mathbf{i}^T\mathbf{U}^{-1}$ and $\mathbf{j}^T\mathbf{U}^{-1}$ such that the skew, r_s, and difference in length, r_a, are given (up to scale) by:

$$r_s = \mathbf{i}^T\mathbf{U}^{-1}\mathbf{U}^{-T}\mathbf{j} \tag{3}$$

$$r_a = \mathbf{i}^T\mathbf{U}^{-1}\mathbf{U}^{-T}\mathbf{i} - \mathbf{j}^T\mathbf{U}^{-1}\mathbf{U}^{-T}\mathbf{j}. \tag{4}$$

Ideally, $\mathbf{i}^T$ and $\mathbf{j}^T$ are orthogonal and have unit aspect ratio such that $r_s = r_a = 0$. For three or more views, there are at least six constraints on $\mathbf{U}^{-1}\mathbf{U}^{-T} = \mathbf{\Omega}^{-1}$ such that r_s and r_a are minimized by a linear least squares solution for $\mathbf{\Omega}^{-1}$.

Little treatment, however, has been presented where there are fewer than three views such that there is insufficient information to upgrade structure and motion without additional constraints (an infinite number of solutions exist for $\mathbf{\Omega}^{-1}$). In the context of human pose estimation, Taylor [16] showed that knowing the *ratios* of lengths and manually solving depth ambiguities was sufficient to recover scene structure, up to a depth scale.

[3] We refer to quantities associated with a particular frame using superscripts within parentheses e.g. $\mathbf{A}^{(f)}$.

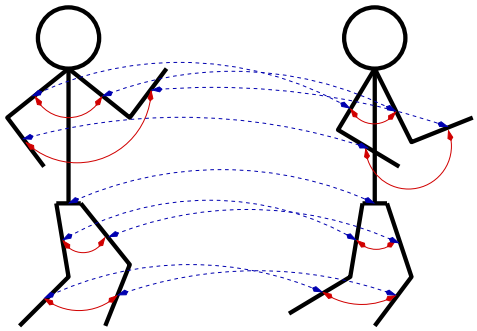

Figure 2: Symmetry (solid) and rigidity (dashed) constraints between a pair of reconstructions.

However, we adopt the approach of Liebowitz and Carlsson [11] which we now review in some detail. In addition to *weakly* constraining motion, they also impose constraints upon known structure such as symmetry and the piecewise rigidity of an articulated body. For a single reconstruction the symmetry of the body provides four constraints between the upper arms, forearms, thighs and forelegs (solid arrows in Fig. 2). Similarly, for any pair of reconstructions the rigidity of the limbs impose nine constraints on each upper arm, forearm, thigh and foreleg, and hips (dashed arrows in Fig. 2).

More formally, the vectors $\mathbf{X}_A$ and $\mathbf{X}_B$ representing *different* links of equal length in the *same* affine reconstruction transform to $\mathbf{U}\mathbf{X}_A$ and $\mathbf{U}\mathbf{X}_B$ where the Euclidean difference in length, r_{sym}, is given by:

$$\begin{aligned} r_{sym} &= \mathbf{X}_A^T\mathbf{U}^T\mathbf{U}\mathbf{X}_A - \mathbf{X}_B^T\mathbf{U}^T\mathbf{U}\mathbf{X}_B \\ &= (\mathbf{X}_A^T - \mathbf{X}_B^T)\boldsymbol{\Omega}(\mathbf{X}_A - \mathbf{X}_B) \end{aligned} \quad (5)$$

Likewise, $\mathbf{X}_A^{(i)}$ and $\mathbf{X}_A^{(j)}$ representing the *same* rigid link in *different* affine reconstructions, i and j, constrain both $\boldsymbol{\Omega}^{(i)}$ and $\boldsymbol{\Omega}^{(j)}$ and the residual error between reconstructions, r_{rig}, is given by:

$$r_{rig} = (\mathbf{X}_A^{(i)})^T\boldsymbol{\Omega}^{(i)}(\mathbf{X}_A^{(i)}) - (\mathbf{X}_A^{(j)})^T\boldsymbol{\Omega}^{(j)}(\mathbf{X}_A^{(j)}). \quad (6)$$

Since rigidity constraints are not independent, Liebowitz and Carlsson apply them between consecutive pairs of reconstructions[4]. Furthermore, since structure and motion constrain $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$, respectively, all residuals cannot be minimized within a linear least squares framework.

Liebowitz and Carlsson [11] address this by optimizing a cost function directly over the $6F-1$ elements of all $\mathbf{U}$, up to a global scale, to recover *local* structure and motion (since rotation and translation between reconstructions is not recovered). This cost function is the sum of squared residuals for both motion and structure. However, these can be weighted arbitrarily to reflect confidence in the constraints.

Since the scale of the structure is fixed by rigidity constraints, the scale of each image is recovered from the computed 'Euclidean' motion. Under the assumption of static cameras, the image measurements of the dynamic scene are then scaled accordingly and treated as an *orthographic* projection of FN features in a *static* scene. A single factorization-rectification operation then recovers a single $\mathbf{P}$ for the entire sequence and *global* structure where rotation and relative translation between frames is recovered (although the coordinate frame remains arbitrary). The structure is then approximated by an articulated body with the median computed segment lengths, in the estimated pose at each frame.

Although theoretically sound, the algorithm presented in [11] has a number of practical limitations such as inefficiency (optimization is performed over $6F-1$ parameters), lack of an intuitive initialization (linear solutions for $\boldsymbol{\Omega}$ are seldom positive definite) and considerable ambiguity in implementation ($\mathbf{U}$ can be parameterized in several different ways, motion and structure costs must be weighted appropriately and several image scales may be selected from the recovered motion).

3. Synchronization extensions

In Section 2.1, we briefly described a synchronization method under the assumption of non-rigid motion viewed with an affine camera. In this section, we show how a rank constraint framework can be applied for the perspective and planar projective (homography) camera models to compute the cost surface, $c(f, f')$. Methods for robustly recovering α and δt from this cost surface are detailed in [18].

3.1. Perspective projection model

Caspi *et al* previously presented a feature-based method for perspective projection [4]. However, they used a geometric distance measure whereas we show that comparable results can be achieved using a computationally cheap *algebraic* distance measure expressed in a rank constraint framework.

Corresponding homogeneous image features, $\mathbf{x}$ and $\mathbf{x}'$, are related by the perspective fundamental matrix[5], $\mathbf{F}$ such that $\mathbf{x}^T\mathbf{F}\mathbf{x}' = 0$. Using the modified 'eight-point' algorithm [8] we compose a matrix, $\mathbf{M}_F$, such that under noiseless conditions $\mathbf{M}_F\mathbf{f} = \mathbf{0}$ where $\mathbf{f} = (f_1, ..., f_9)^T$ is the normalized vector of non-zero elements of $\mathbf{F}$, recovered using the SVD of $\mathbf{M}_F$. We define a match cost, $c(f, f')$, as the sum of squared *algebraic* distances, $d_{alg}(\cdot)$, between the points $\mathbf{x}_i$ and their epipolar lines, $\mathbf{F}\mathbf{x}'_i$ such that:

[4] Experiments suggest they should be applied with respect to the *same* reconstruction to avoid drift in the scale of the reconstructions.

[5] Using the affine fundamental matrix results in the affine method described in Section 2.1.

$$c(f, f') = \sum_i d_{alg}(\mathbf{x}_i, \mathbf{F}\mathbf{x'}_i)^2 = \|\mathbf{M}_F\mathbf{f}\|^2 = \sigma_9^2 \qquad (7)$$

where σ_i is the ith singular value of $\mathbf{M}_F$.

This was tested on a running sequence pair (Fig. 3), offset by 30 frames, where we manually labelled both points on the body and points on the ground plane such that perspective effects were observable. Using the known epipolar geometry of the cameras, target image features were projected onto their corresponding epipolar lines to reduce tracking error. The algorithm recovered an offset of $\delta t = 30$ (since the features were noiseless). In comparison, the affine method [18] recovered an inaccurate estimate of $\delta t = 29.69$ since the projection model was inappropriate given the depth of the scene.

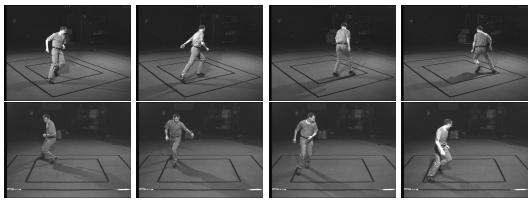

Figure 3: Running sequence seen from the (top) reference and (bottom) target viewpoints.

3.2. Homography model

Synchronizing sequences related by a homography was also studied by Caspi and Irani using optical flow [2, 3]. We consider the case where we observe point features moving independently in a plane.

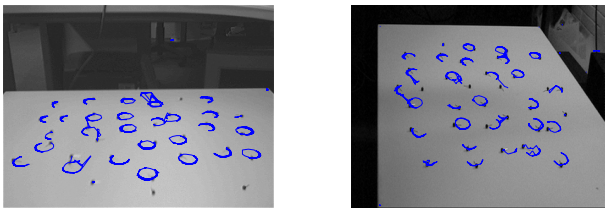

Figure 4: Frames from the homography sequences. The feature tracks are elliptic due to the circular motion of the points although this does not influence the generality of the method.

Corresponding homogeneous image features, $\mathbf{x}$ and $\mathbf{x}'$, are related by a homography, $\mathbf{H}$, such that $\mathbf{H}\mathbf{x} = \mathbf{x}' \Rightarrow [\mathbf{x}'_\times]\mathbf{H}\mathbf{x} = [\mathbf{x}'_\times]\mathbf{x}' = \mathbf{0}$ where $[\mathbf{x}'_\times]$ represents the matrix equivalent of the vector product. In a similar manner to perspective projection we define a matrix, $\mathbf{M}_H$, such that under noiseless conditions $\mathbf{M}_H\mathbf{h} = \mathbf{0}$ where $\mathbf{h} = (h_1, ..., h_9)^T$ is the normalized vector of elements of $\mathbf{H}$. Again, we define $c(f, f')$ as the squared algebraic distances between measurements $\mathbf{x'}_i$ in the second sequence and transferred features, $\mathbf{H}\mathbf{x}_i$, from the first sequence:

$$c(f, f') = \sum_i d_{alg}(\mathbf{x'}_i, \mathbf{H}\mathbf{x}_i)^2 = \|\mathbf{M}_H\mathbf{h}\|^2 = \sigma_9^2 \qquad (8)$$

where σ_i is the ith singular value of $\mathbf{M}_H$. Again, recovering $\mathbf{H}$ from $\mathbf{M}_H$ using the SVD is trivial.

We demonstrate this method with point features moving independently in a plane, captured using two cameras at approximately 12.5 frames/sec and 8 frames/sec. A crude feature tracker was implemented to recover the feature tracks (Fig. 4). Although many tracks were corrupted by noise and tracking error, thirteen were selected and matched by hand. The true parameter values were manually estimated as $\alpha \approx 0.64$ and $\delta t \approx 16$, which correspond closely to the recovered values of $\alpha = 0.6249$ and $\delta t = 14.04$.

4. Improved self-calibration

We now describe an improved self-calibration method (Algorithm 1) that exploits a minimal parameterization of $\mathbf{\Omega}$ based on assumptions regarding the camera calibration. Specifically, we *assume* the cameras have zero skew and unit aspect ratio such that we need only minimize structural error over all $\mathbf{\Omega}$ that *exactly* satisfy these constraints. This is followed by a full bundle adjustment over the free parameters to minimize a geometric reprojection error.

I. Recover all $\mathbf{U}$ using a minimal parameterization of $\mathbf{\Omega}$ and compute local structure and motion;

II. Normalize image feature locations and process again to compute global structure and motion;

III. Perform bundle adjustment over all free parameters, minimizing geometric reprojection error.

Algorithm 1: Novel approach.

4.1. Minimal parameterization

By strictly enforcing motion constraints, we eliminate four degrees of freedom in $\mathbf{\Omega}^{-1}$. Combining the four bilinear motion constraints in six unknowns results in a matrix whose nullspace is spanned by two possible values for $\mathbf{\Omega}^{-1}$ (denoted by $\mathbf{\Omega}_1^{-1}$ and $\mathbf{\Omega}_2^{-1}$) any linear combination of which satisfies all motion constraints *exactly*. We therefore parameterize all such $\mathbf{\Omega}^{-1}$ using polar coordinates:

$$\mathbf{\Omega}^{-1}(r, \theta) = r(\cos(\theta) \cdot \mathbf{\Omega}_1^{-1} + \sin(\theta) \cdot \mathbf{\Omega}_2^{-1}) \qquad (9)$$

although this does not force $\mathbf{\Omega}^{-1}$ to be positive definite. From the known composition of $\mathbf{\Omega}^{-1}$, we compute the range

$(\theta_{min}, \theta_{max})$ for which $\mathbf{\Omega}^{-1}$ is positive definite and minimize the cost over all $r > 0$ and $\theta_{min} < \theta < \theta_{max}$ such that all $\mathbf{U}$ can be recovered by Cholesky factorization.

The cost increases to infinity at θ_{min} and θ_{max} and is convex in between such that unconstrained methods can be employed and the midpoint of this interval provides an intuitive initialization for optimization. Furthermore, the total number of parameters is reduced from $6F - 1$ to $2F - 1$, implementation is unambiguous and no parameters are required to be tuned empirically.

4.2. Bundle adjustment

Having recovered local (Stage I) and then global (Stage II) structure, we recover the median body segment lengths[6] and pose. The recovered structure at each frame is then replaced with an articulated model at the estimated pose as in [11].

We then minimize the geometric reprojection error using an affine bundle adjustment to optimize over all free parameters: image scales, s_i^f; camera rotation, $(\theta_x, \theta_y, \theta_z)$; camera translation, $\mathbf{t}$; body segment lengths, $\mathbf{L}$; and pose parameters, $\boldsymbol{\phi}^f$. We assume the cameras have unit aspect ratio and zero skew. Defining $\boldsymbol{\epsilon}$ as the vector of reprojection errors over all measurements, we seek to minimize the sum of squared reprojection errors, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, over all F frames:

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum_i \sum_f \sum_n \|\mathbf{x}^f_{i,\,n} - (s_i^f \mathbf{R}\mathbf{X}_n(\mathbf{L}, \boldsymbol{\phi}^f) + \mathbf{t})\|^2_F \quad (10)$$

where $\mathbf{R}$ is a rotation matrix computed using Euler angles $(\theta_x, \theta_y, \theta_z)$, and $\mathbf{X}_n(\mathbf{L}, \boldsymbol{\phi})$ is the position of the nth feature given link lengths, $\mathbf{L}$, and pose, $\boldsymbol{\phi}$. The minimization is achieved by solving the normal equations $\mathbf{J}^T\mathbf{J}\boldsymbol{\Delta} = \mathbf{J}^T\boldsymbol{\epsilon}$ for $\boldsymbol{\Delta}$ where $\mathbf{J}$ is the Jacobian of all measurements with respect to the parameters. Since pose and scale are frame dependent, $\mathbf{J}$ is sparse permitting efficient minimization. The result is an articulated model fitted to the subject (up to scale), capturing the pose at every frame. For comparison, we also implement a perspective bundle adjustment, initialized using conventional calibration techniques[7].

4.3. Results

Two 30 frame image sequences of a running motion were synthesized using motion capture data from a commercial system. An articulated model of known segment lengths was imaged under perspective projection and the projected image features used to recover affine structure by factorization. Euclidean structure and motion was then recovered using four methods: (i) Liebowitz and Carlsson (L&C) rectification; minimal parameter rectification with (ii) no bundle adjustment; (iii) affine bundle adjustment; (iv) perspective bundle adjustment.

Table 1 shows a comparison of the methods for noiseless data based upon (i) number of iterations required for convergence of minimization routines, (ii) time taken for convergence[8], (iii) total time taken (including fixed overhead costs) and (iv) final RMS reprojection error, E_{RMS} (pixels).

		L&C	Minimal	A.B.A	P.B.A
I	# iter.	16	10	10	10
	Time (s)	2.08	0.68	0.62	0.62
II	# iter.	382	6	6	6
	Time (s)	2.84	0.058	0.058	0.058
III	# iter.	-	-	12	85
	Time (s)	-	-	16.25	160.81
Total time (s)		6.96	2.81	23.00	233.4
E_{RMS}		1.41	1.44	0.785	2.9×10^{-4}

Table 1: Performance comparison of four methods.

Table 2 shows how performance degrades with added isotropic, zero-mean Gaussian noise of increasing standard deviation, σ. Comparisons are based upon the mean percentage error in recovered segment length, E_L, mean RMS error in joint angle, E_J, over the knee and elbow joints and camera rotation error.

	σ	L&C	Minimal	A.B.A	P.B.A
E_L	0	0.871	0.905	0.724	0.001
	2	7.830	6.195	2.561	2.415
	4	10.10	10.60	8.666	8.256
E_J	0	0.0521	0.0511	0.0328	3.8×10^{-5}
	2	0.2851	0.2776	0.1712	0.1644
	4	0.3390	0.3435	0.3255	0.3220
E_ω	0	0.087	0.086	0.048	3.3×10^{-5}
	2	0.317	0.285	0.038	0.006
	4	0.394	0.470	1.8×10^{-4}	0.045
$E_\mathbf{a}$	0	0.101	0.102	0.076	1.79×10^{-5}
	2	0.071	0.076	0.076	0.004
	4	0.055	0.045	0.071	0.010

Table 2: Recovered percentage error in segment lengths, E_J, RMS error in joint angles, E_J (rad), and error in camera rotations, E_ω (rad) and $E_\mathbf{a}$ (rad), with increasing image noise, σ (pixels).

To compare camera rotations we adopt the angle-axis notation. A rotation is represented by a twist angle, ω, about an axis parallel to the unit vector, $\mathbf{a}$. We denote ground truth values by ω_{gt} and $\mathbf{a}_{gt}$, quantifying error using the difference in twist angle, $E_\omega = |\omega_{gt} - \omega|$ and the angle between the axes, $E_\mathbf{a} = \cos^{-1}(\mathbf{a}_{gt}^T\mathbf{a})$.

Our results demonstrate that the accuracy of the minimal parameterization is comparable to that of [11] whilst reduc-

[6] We no longer enforce symmetry at this stage since this is considered to be the most uncertain of our assumptions.

[7] Due to instability, we use a stratified bundle adjustment. See [18] for details.

[8] Based on a 2.4GHz Pentium 4 desktop computer

ing computational complexity and eliminating implementation ambiguity. Furthermore, affine bundle adjustment generally provides a noticeable improvement in accuracy - often comparable to that of perspective bundle adjustment (theoretically the best we can hope to achieve in a Maximum Likelihood framework).

However, we observe a sharp increase in E_L with image noise, even for perspective bundle adjustment since this error is dependent on the imaged size of the subject and even small amounts of noise may incur a large percentage error in projected length. This is exacerbated further with out-of-plane rotation due to foreshortening effects.

5. Motion capture of a novel sequence

Finally, we demonstrate the complete system combining the synchronization and self-calibration methods described in this paper. A juggling sequence was captured from two different viewpoints (Fig. 5) using uncalibrated and unsynchronized cameras of different frame rates.

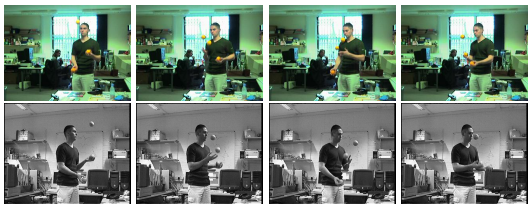

Figure 5: Corresponding frames from the juggling sequence

Specifically, the reference sequence was captured using an NTSC digital camera and consisted of 150 colour frames at 30Hz with a resolution of 320×240 pixels whilst the target sequence, captured with a PAL analogue camera, contained 250 greyscale frames at 25Hz with a resolution of 720×576 pixels. Corresponding feature locations on the upper body, head and juggling balls were marked manually but were not corrected since the epipolar geometry of the cameras was unknown.

From the known frame rates, we computed $\alpha = 25/30 \approx 0.833$ and estimated that $\delta t \approx 115$ by inspection. These estimates were in close agreement with the recovered values of $\alpha = 0.8371$ and $\delta t = 113.60$. New target feature locations corresponding to the reference locations were synthesized using linear interpolation of the locations in the original target sequence. Affine structure and motion was then computed by factorization and upgraded to a Euclidean coordinate frame (Fig. 6) using the self-calibration method described in Section 4 with affine bundle adjustment.

Table 3 shows the recovered body segment lengths where we see that the symmetry has been recovered and the segments are in proportion, despite the reduced number of structural constraints (the lengths are normalized with respect to the upper left arm). Finally, Fig. 7 shows the joint trajectories of the elbows during the motion where the periodic motion is apparent in addition to the expected phase difference.

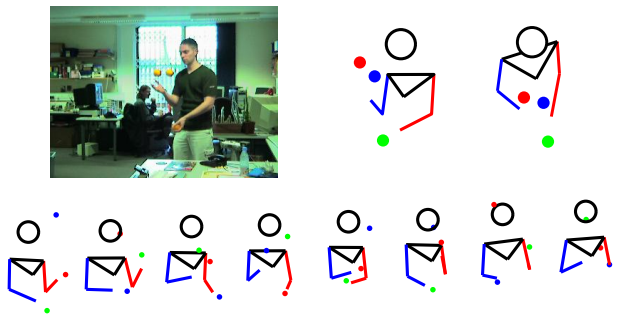

Figure 6: (top left) Frame 75 (top centre) reconstruction from camera 1 (top right) reconstruction from novel view (bottom) Sequence of reconstructions from novel view.

Limb	Left	Right
Upper arm	1.000	1.032
Lower arm	0.984	0.982

Table 3: Recovered limb lengths (relative to the left upper arm) for the juggling sequence

6. Conclusion

We have presented a method of recovering non-rigid metric structure and motion from two unsynchronized and uncalibrated cameras by combining synchronization and self-calibration techniques, demonstrated for the application of human motion capture.

We extend current methods by applying rank constraint based synchronization algorithms for perspective projection and homography models whilst exploiting a minimal parameterization of the rectifying affine transformation to upgrade affine structure to a Euclidean coordinate frame. This parameterization provides an unambiguous implementation (requiring no parameters to be tuned empirically), reduces computational complexity and provides an intuitive initialization for optimization. Pose recovery is then completed using a full bundle adjustment over the free parameters.

Among areas requiring further development, rectification is strictly a batch process so an obvious extension would be to develop a recursive method. However, the key requirement of the system is that of spatial feature correspondence, solved by manual labelling (or markers). As a result, the principal direction of future development will be in the recovery of joint locations using image data only.

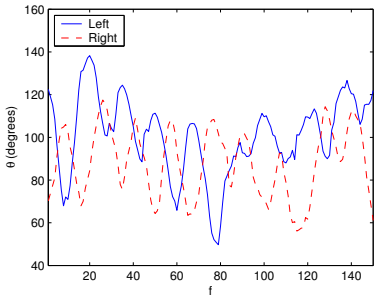

Figure 7: Recovered trajectories of the elbows during juggling

Acknowledgments

This project was funded by the Engineering and Physical Sciences Research Council (EPSRC). Image sequences were provided by Oxford Metrics. Data used in this project was obtained from mocap.cs.cmu.edu, created with funding from NSF EIA-0196217.

References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. 17th IEEE Conf. on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 8–15, June 1998.

[2] Y. Caspi and M. Irani. A step towards sequence-to-sequence alignment. In *Proc. 19th IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island*, volume 2, pages 682–689, June 2000.

[3] Y. Caspi and M. Irani. Alignment of non-overlapping sequences. In *Proc. 8th Int'l Conf. on Computer Vision, Vancouver*, volume 2, pages 76–83, July 2001.

[4] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. In *Proc. Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen*, May 2002.

[5] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *Proc. 5th Int'l Conf. on Computer Vision, Boston*, pages 1071–1076, June 1995.

[6] J. Deutscher, A. Blake, and I. D. Reid. Markerless motion capture of complex full-body movement for character animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[7] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *Proc. 15th IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, California*, pages 73–80, June 1996.

[8] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, June 1997.

[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[10] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1453–1459, December 2000.

[11] D. Liebowitz and S. Carlsson. Uncalibrated motion capture exploiting articulated structure constraints. In *Proc. 8th Int'l Conf. on Computer Vision, Vancouver*, volume 2, pages 230–237, July 2001.

[12] D. W. Pooley, M. J. Brooks, A. J. van den Hengel, and W. Chojnacki. A voting scheme for estimating the synchrony of moving-camera videos. In *Proc. International Conference on Image Processing, Barcelona, Spain*, September 2003.

[13] L. Quan. Self-calibration of an affine camera from multiple views. *International Journal of Computer Vision*, 19(1):93–110, 1996.

[14] I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *International Journal of Computer Vision*, 18(1):41–60, April 1996.

[15] I. D. Reid and A. Zisserman. Goal-directed video metrology. In B. Buxton and R. Cipolla, editors, *Proc. 4th European Conf. on Computer Vision, Cambridge*, volume 2, pages 647–658, April 1996.

[16] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Proc. 19th IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island*, volume 1, pages 677–684, June 2000.

[17] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[18] P. Tresadern and I. Reid. Human motion capture from two uncalibrated and unsynchronized cameras. *submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[19] P. Tresadern and I. Reid. Synchronizing image sequences of non-rigid objects. In *Proc. 14th British Machine Vision Conf., Norwich*, volume 2, pages 629–638, September 2003.

[20] Vicon Motion Systems. VMS009S_600_TechSheet.pdf. Available at http://www.vicon.com/engineering/downloads/.

[21] L. Wolf and A. Zomet. Correspondence-free synchronization and reconstruction in a non-rigid scene. In *Proc. Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen*, May 2002.

[22] L. Zelnik-Manor and M. Irani. Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations. In *Proc. 21st IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 287–293, June 2003.

[23] C. Zhou and H. Tao. Dynamic depth recovery from unsynchronized video streams. In *Proc. 21st IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 351–358, June 2003.